www.gssrjournal.com
Global
Social Sciences Review
exploring humanity

p-ISSN : 2520-0348 | e-ISSN : 2616-793X

DOI(Journal): 10.31703/gssr
DOI(Volume): 10.31703/gssr/.2024(IX)
DOI(Issue): 10.31703/gssr.2024(IX.I)

GSSR

GLOBAL SOCIAL SCIENCES REVIEW

HEC-RECOGNIZED CATEGORY-Y

VOL. IX, ISSUE I, WINTER (MARCH-2024)

Double-blind Peer-review Research Journal
www.gssrjournal.com

HumanityPublications
sharing research
www.humapub.com
US | UK | Pakistan

Humanity Publications(HumaPub)
www.humapub.com
Doi:https://dx.doi.org/10.31703

Article Title

Automatic Spoofing Detection Using Deep Learning

Global Social Sciences Review

p-ISSN: 2520-0348**e-ISSN:** 2520-0348

DOI(journal): 10.31703/gssr

Volume: IX (2024)

DOI (volume): 10.31703/gssr.2024(IX)

Issue: I (Winter-March 2024)

DOI(Issue): 10.31703/gssr.2024(IX-I)

Home Page
www.gssrjournal.com

Volume: IX (2024)
https://www.gssrjournal.com/Current-issues

Issue: I-Winter (March-2024)
https://www.gssrjournal.com/Current-issues/9/1/20234

Scope
https://www.gssrjournal.com/about-us/scope

Submission
https://humaglobe.com/index.php/gssr/submissions

Google Schlor

Visit Us

Abstract

Deep fakes stand out to be the most dangerous side effects of Artificial Intelligence. AI assists to produce voice cloning of any entity which is very arduous to categorize whether it's fake or real. The aim of the research is to impart a spoofing detection system to an automatic speaker verification (ASV) system that can perceive false voices efficiently. The goal is to perceive the unapparent audio elements with maximum precision and to develop a model that is proficient in automatically extracting audio features by utilizing the ASVspoof 2019 dataset. Hence, the proposed ML-DL SafetyNet model is designed that delicately differentiate ASVspoof 2019 dataset voice speeches into fake or bonafide. ASVspoof 2019 dataset is characterized into two segments LA and PA. The ML-DL SafetyNet model is centred on two unique processes; deep learning and machine learning classifiers. Both techniques executed strong performance by achieving an accuracy of 90%.

Keywords: Fake Audio, Spoof Speech Detection, Deep Learning

Authors:

Muhammad Nafees: (Corresponding Author)
MSc, Department of Data Science, University of Engineering and Technology, Taxila, Punjab, Pakistan.
(Email: engrmnafees512@gmail.com)

Abid Rauf: MSc, Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.

Rabbia Mahum: MS, Department of Computer Science, University of Engineering and Technology, Taxila, Punjab, Pakistan.

Pages: 111-133

DOI:10.31703/gssr.2024(IX-I).11

DOI link:https://dx.doi.org/10.31703/gssr.2024(IX-I).11

Article link: http://www.gssrjournal.com/article/A-b-c

Full-text Link: https://gssrjournal.com/fulltext/

Pdf link: https://www.gssrjournal.com/jadmin/Auther/31rvIolA2.pdf

Citing this Article

11		Automatic Spoofing Detection Using Deep Learning						
		Author	Muhammad Nafees Abid Rauf RabbiaMahum		DOI	10.31703/gssr.2024(IX-I).11		
Pages		111-133	Year	2024	Volume	IX	Issue	I
Referencing & Citing Styles	APA 7th	Nafees, M., Rauf, A., & RabbiaMahum. (2024). Automatic Spoofing Detection Using Deep Learning. *Global Social Sciences Review*, *IX*(I), 111-133. https://doi.org/10.31703/gssr.2024(IX-I).11						
	CHICAGO	Nafees, Muhammad, Abid Rauf, and RabbiaMahum. 2024. "Automatic Spoofing Detection Using Deep Learning." *Global Social Sciences Review* IX (I):111-133. doi: 10.31703/gssr.2024(IX-I).11.						
	HARVARD	NAFEES, M., RAUF, A. & RABBIAMAHUM 2024. Automatic Spoofing Detection Using Deep Learning. *Global Social Sciences Review,* IX, 111-133.						
	MHRA	Nafees, Muhammad, Abid Rauf, and RabbiaMahum. 2024. 'Automatic Spoofing Detection Using Deep Learning', *Global Social Sciences Review*, IX: 111-33.						
	MLA	Nafees, Muhammad, Abid Rauf, and RabbiaMahum. "Automatic Spoofing Detection Using Deep Learning." *Global Social Sciences Review* IX.I (2024): 111-33. Print.						
	OXFORD	Nafees, Muhammad, Rauf, Abid, and RabbiaMahum (2024), 'Automatic Spoofing Detection Using Deep Learning', *Global Social Sciences Review,* IX (I), 111-33.						
	TURABIAN	Nafees, Muhammad, Abid Rauf, and RabbiaMahum. "Automatic Spoofing Detection Using Deep Learning." *Global Social Sciences Review* IX, no. I (2024): 111-33. https://dx.doi.org/10.31703/gssr.2024(IX-I).11.						

e-ISSN:2520-0348 **Volume:** IX (2024) **Issue:** I-Winter (March-2024) **p-ISSN**:2520-0348

Global Social Sciences Review

www.gssrjournal.com

DOI: http://dx.doi.org/10.31703/gssr

Pages: 111-133 **URL:** https://doi.org/10.31703/gssr.2024(IX-I).11 **Doi:** 10.31703/gssr.2024(IX-I).11

Title

Automatic Spoofing Detection Using Deep Learning

Authors:

Muhammad Nafees: (Corresponding Author)
MSc, Department of Data Science, University of Engineering and Technology, Taxila, Punjab, Pakistan.
(Email: engrmnafees512@gmail.com)

Abid Rauf: MSc, Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.

RabbiaMahum: MS, Department of Computer Science, University of Engineering and Technology, Taxila, Punjab, Pakistan.

Contents

Abstract

Deep fakes stand out to be the most dangerous side effects of Artificial Intelligence. AI assists to produce voice cloning of any entity which is very arduous to categorize whether it's fake or real. The aim of the research is to impart a spoofing detection system to an automatic speaker verification (ASV) system that can perceive false voices efficiently. The goal is to perceive the unapparent audio elements with maximum precision and to develop a model that is proficient in automatically extracting audio features by utilizing the ASVspoof 2019 dataset. Hence, the proposed ML-DL SafetyNet model is designed that delicately differentiate ASVspoof 2019 dataset voice speeches into fake or bonafide. ASVspoof 2019 dataset is characterized into two segments LA and PA. The ML-DL SafetyNet model is centred on two unique processes; deep learning and machine learning classifiers. Both techniques executed strong performance by achieving an accuracy of 90%.

Keywords: Fake Audio, Spoof Speech Detection, Deep Learning

Introduction

As per researchers and UCL report conclusions, fake audio has become a serious and challenging problem nowadays and appears to be one of the most perturbing applications of artificial intelligence. Likewise, research also indicates that artificial intelligence will assist criminality in different tactics for the next 15 years. Furthermore, the architectures exercised by deep fakes are so concentrated and complex, that it is gruelling to segregate and prevent authentic and fake speeches.

In the recent era, there has been a sudden growth in the technology for speaker verification systems. But as far as its useful and beneficial aspects, instead, some serious risks exist there regarding improvement in the technology. Replay spoof attacks [1]can be easily operated with smartphones by utilizing AI tools, as they do not require any prior professional experience. Furthermore, it is quite challenging for the ASV (Automatic Speaker Verification) systems, as it is difficult to identify the authentic and fake speeches. The whole audio is altered by the same voice as the genuine speaker in audio-deep fake attacks, [Gao2021]propounding a considerable threat**.** For example, a hacker may get control over private information and contents by effectively developing fraudulent voices to encrypt voice-print-based security systems. Likewise, a person can also deceive a bank call centre by identifying himself as a registered user and may successfully make bank representatives transfer money to his account. Furthermore, an attacker can get access to the security system that is established on the voiceprint. With the growth and advancement in technology, it is quite challenging to handle this issue artificial intelligence technology has become advanced now. As for AI technology's negative impacts, on the contrary, AI technology assisted the researchers in contributing to protection against deep fake problems, like machine learning and deep learning tools. If we look a few years back, there is a considerable advancement in the field of artificial intelligence and strong architectures are utilized that even humans cannot distinguish genuine and fake speeches. These technologies [Yi2022], [Kinnunen2017], [Todisco2019] can be used for criminal activities or illegal activities and these technologies have the capacity to affect the credibility of frequently employed biometric identification models. It is necessary to address and develop strategies for recognizing the serious damage that false audio can cause. In the same way, artificial intelligence has a great role in the field of forensics. The research is being carried out for development in the field of forensics. The objective of the research [Mcuba2022]is to detect fake audio and their association by using various deep learning techniques so that deep fakes are identified at the earliest stage. The model performed various techniques of deep learning like Mel-spectrum, MFCC etc. to accomplish improved results. Furthermore, among these techniques architecture of VGG-18 performed best for finding real and fake speeches for forensics. The importance, role and requirement of machine learning and deep learning in current years are amplified with the evolution in technology. The research [Hamza2022] confers the technique of MFCC being carried out to achieve the information regarding the audios being fake or bonafide. As the modern approaches for deep fakes are so effective that it is very hard to categorize these attacks. The genuine and fake dataset was selected and then parted into different four datasets as per the investigation. The results of the findings [Almutairi 2023] obviously show that among the various machine learning algorithms SVM is the best for chosen dataset. The research depicts the method for automatically detecting the fake audios regarding Arabic speech, as limited work is carried out for distinctive languages spoofing detection. A dataset is created based upon the modern speech of Arabic pronunciation and speech was then tested and trained with the people who are non-Arabic speakers. By using their own model for spoofing detection, the researchers achieved a good accuracy based on EER. Moreover, our proposed research could be considered to be interdisciplinary as audios are part of linguistics.

Related Work

Several research studies have been published that deal with automatic spoofing detection. MissimilianoTodisco's model shows the importance of developing techniques against threats of genuine and false audios. Jiyangyan Yi developed [Yi 2021]a dataset for half-truth audio detection. MoustafaAlzantot proposed model [Alzantot2019]goal is to discriminate genuine and spoofing speeches by establishing strong defensive structures. Galina Lavrentyevastruggles [Lavrentyeva 2017] to perceive the spoofing by using a deep learning approach for ASV spoof 2017 by using anti spoofing system. The proposed model for the ASV spoof 2017 challenge succeeded with an accuracy of 87% by using a mixture of CNN, SVM, CNN and RNN networks. B.T Balamuralai proposed [Balamurali 2019]classical GMM-UBM model achieved the comparative results of mixed machine learning and identified audio structures. Shanshan Zhang proposed [Zhang 2022] model aimed to distinguish false speeches by using pre-training models. The investigation [Dua2022] attempts to detect speaker verification by using deep learning models. The findings of the research are the fusion of two models having time-distributed dense layers, LSTM and deep neural networks. The fusion model in this investigation performed well for CQCC features.

The research [15] is based on spoof speech detection for fake speech, voice alteration and replay attack techniques. The dawn of this era [Wenger2021] has introduced various tools that are used to deceive the world by producing audios and speeches that sound authentic as spoken by the target speaker. Furthermore, in case if these tools fall into the wrong hands, will create great hazards for the world. The hazard can be at the personal level, organizational level or at the world level. The research actually highlights and efforts the impacts of these tools on machines and computers. Findings in the research clearly show that machines and humans can easily be fooled by the latest tools and techniques. Therefore, it is suggested to raise awareness and to develop the latest and advanced protections to protect our machines, systems and humans. As per the researcher's interest studies [Tan 2021], text-to-speech is the most trending topic in the field of artificial intelligence which has a variety of applications. Currently, deep learning technology has improved these TTS techniques. The survey research is based on TTS which highlights the current research in the field of deep learning by utilizing various relations. The research [18] presents a corpus named SAS which comprises nine techniques of spoofing of which two are dialogue fusion and the remaining are speech renovation. Research is designed for two protocols each performing a different duty. One protocol is used for evaluating speaker verification and the other protocol is used for creating spoofing material. The research is based on the utilization of recent ideas and in the absence of any verbal language (audio) spoofing detection technique, the system has a greater chance of being attacked. This paper [Columbia2021] highlights serious threats related to spoofing. For this purpose, main ambition of researchers is to find out the spoofed speech from the bonafide one. In this research a method is utilized named the capsule network by using ASV spoof 2019 dataset for detecting audios. Major work done is based on text to speech conversion. Moreover, replay attacks were also taken as part of research and results clearly show that the model also performed well in this case.

Communication networks [15] are taken under consideration by using MAC addresses of different devices. The main objective of the research was to detect the MAC address in a wireless medium. In this research, the experiment was conducted by using different distances from devices. This system does not depend upon the amendments of standards and protocols of the devices. The system achieved different results on the basis of the targeted device distance. The model performed best for random forests. The recent world is progressing as fast as the speed of light. Many technologies [20] have been introduced in the arena of computer science. Machine learning has also played a vital role. Nowadays various machine learning tools can be used to automatically create different deep fakes which are very similar to the real ones. Currently, deep fake videos have created a condition of distress in the world. Using ML tools, not only the common public is threatened but celebrities, politicians, actors and many other high-profile persons have been threatened. For this purpose, a model is suggested based on the CNN which is used to automatically detect the spoofed videos. RNN technique is used in the paper which differentiates the spoofed and the bonafide videos. The dataset for this research has been gathered from the various sites. As the advancement in technology [21] is growing frequently and people are facilitated by means of this technology, their worries and security concerns are also rising. Smartphone is one of the major technologies of this era. We use various apps on our smartphones and face movement, and mouth movement features are utilized in this model for detection. The application MoviePy is utilized in which cutting and editing are done on the image data containing the mouth exposed along with the visibility of teeth. DFT and CNN techniques are used to achieve the results for the detection of fake and real videos. With the passage of time and advancements in artificial intelligence techniques Ismail2021], public privacy and security are at high risk. People use AI complex architecture techniques to threaten the public by creating various fake videos in which the face of someone else is being swapped with the targeted person. Face-swapping detection is a challenging task to identify whether the video is false or genuine. The research proposed model YOLO face detector is utilized and ResNet CNN is used to excerpt structures from video frames after getting these structures XGBOOST identify either the video is fake or real.

Table 1

Summary of Some of the Papers Related to the Proposed Network

S. No	Authors	Dataset used	Technique	Accuracy	Future Work
1	Massimiliano Todisco, Xin Wang and Ville Vestman (2019)	ASV spoof 2019	Tendon detection cost function (t-DCF), Gaussian Mixture Model (GMM), constant Q cepstral coefficient	EER 3.92%	NA
2	Jiangyan Yi and Ye Bai (2021)	AISHELL-3 corpus	Gaussian Mixture Model (GMM), Light Convolution Neural Network (LCNN)	82%	Different types of fakes and to develop datasets for other languages
3	MoustafaAlzantot and Ziqi Wang (2019)	78 human voice clips	Linear Frequency Cepstral Coefficients (LFCC), Gaussian mixture models (GMMs)	t-DCF 0.1569% EER 6.02 %	Improving model against unknown attacks
4	Galina Lavrentyeva1, Sergey Novoselov (2017)	ASV spoof 2017	SVM i-vector, LCNN, CNN+RNN	85%	NA
5	B. T. Balamur Ali, Kin Wah Edward Lin (2019)	ASV spoof 2017	MFCCs for audio preprocessing and audio feature selection, and CCs and Autoencoders are used for input and output matching	EER 12.6 %	The proposed architecture can be used with the assistance of the GMM model
6	Mohit Dua, Chhavi Jain (2022)	ASV spoof 2015	LSTM & CNN	1.7% ERR	ASV spoof 2019 dataset to be used for better outputs
7	YanminQiana, Nanxin Chena, Kai Yua, (2019)	ASV spoof 2015	MFCC & PLP are used as deep features. SGD is used to train the constraints and Square Error is used as a function. LSTM & BLSTM are based models, GMM is used to model the input structures while MAP is used to explain the initial model that denotes genuine and spoofed speeches	84% for spoofing discriminant - DNN, 97% for LSTM, 97.2 % for BLSTM	Need to improve EER
8	Emily Wenger, Max Bronckers (2021)	ASV spoof 2017	SVM, Light CNN & Custom CNNs are used with various functions to achieve the results	88%	Exploring subsequent challenges and opportunities
9	Abhijit Jadhav, Abhishek Patange, Patil, (2022)	YouTube, FaceForensics++	ResNet CNN classifier for mining the features. LSTM for Sequence Processing	94%	detection of the audio deep fakes

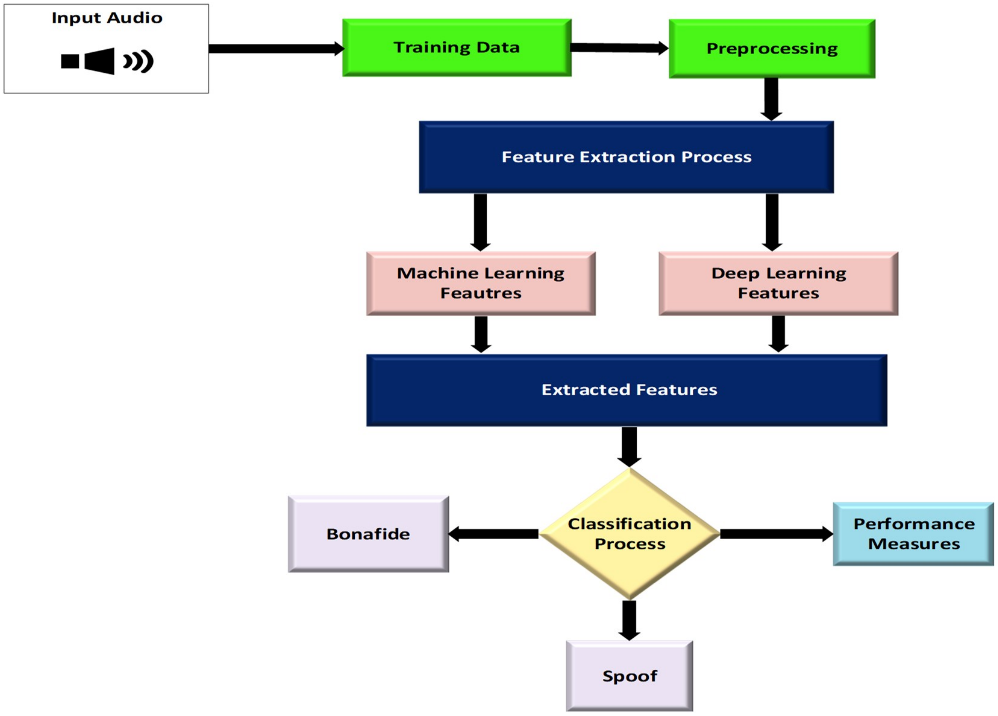
S. No	Authors	Dataset used	Technique	Accuracy	Future Work
10	Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen (2021)	deep fake dataset, FaceForensics dataset	capsule network	95.93%	assessing the capability of the proposed method to fight argumentative machine attacks
11	Oscar de Lima, Sean Franklin, Annet George (2020)	Celeb-DF dataset containing 590 real videos from YouTube and 5639 fake videos	DFT, RCN, R3D, ResNet Mixed 3D-2D Convolution	RCN- 76.25 R2Plus-98.07 I3D- 92.28 MC3- 97.49 R3D- 98.26	–

Methodology

The proposed methodology is divided into categories like data gathering, training and testing. We have collected data from the ASV spoof 2019 corpus containing logical access (LA) and physical access (PA) speeches. The proposed model ML-DL SafetyNet introduced a model, basically a deep learning model and a machine learning model using different features. For the machine learning based, ML-DL SafetyNet model we used different algorithms like Naïve Bayes, and KNN, likewise, using the deep learning ML-DL SafetyNet model we have utilized features of convolutional neural networks and different optimizers and filters like sigmoid, ReLU etc. In the final stage, ML-DL SafetyNet classified the desired outputs. The classifier differentiated the data into the desired categories like spoof and bonafide. The complete methodology process is shown in the figure 1 below.

Figure 1

Complete Flow Chart Detail of the Proposed ML-DL SafetyNet Model



Deep Learning

A deep learning model contains various layers like input layers, hidden layers and classifications layers as shown in the below figure 2. Hidden layers are subcategorized into pooling, batch normalization, convolutional, activation and other different layers. In this model, features are to be extracted using various filters by convolution. When the convolution process is performed, a filter map is generated and dimensions are reduced with the help of pooling to avoid the computational power.

Input Layer

The first layer receives artificial neurons and then transfers these neurons and information to the other networks connected to it.

Hidden Layer

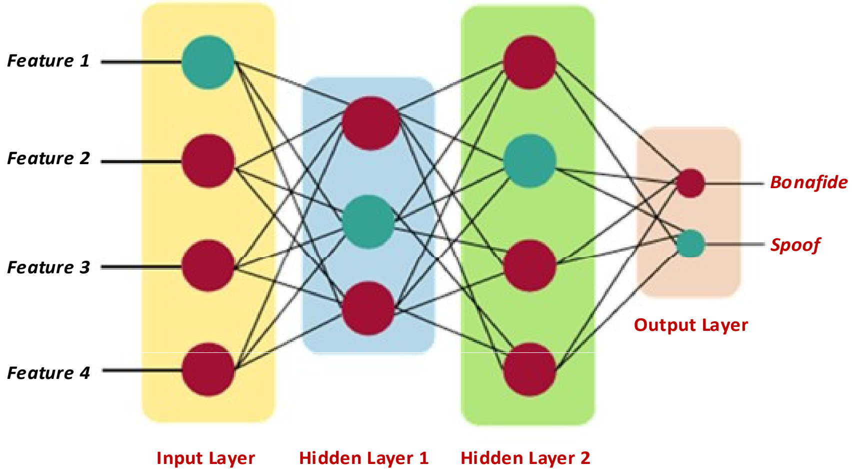
The layers that come after the input layers are hidden layers. These layers vary in their number for a network as per data problem. These layers can be called the backbone of the neural network because these layers are ultimately responsible for the exceptional performance of the network. These are multi-tasking layers, performing different activities and functions at the same time. The hidden layer involved many different layers segregated by their distinct names as per their roles and tasks executed by these layers. Some of these layers are named ReLU, MaxPooling, Fully connected layer, Sigmoid etc. These layers vary in their number depending on the complexity of the network and the computational cost of the system.

Output Layer

This is the final layer which highlights and illustrates the desired predictions of the network. This is the sole layer in the whole network which is tasked to provide the conclusive result.

Figure 2

Detailed Architecture of the Neural Network

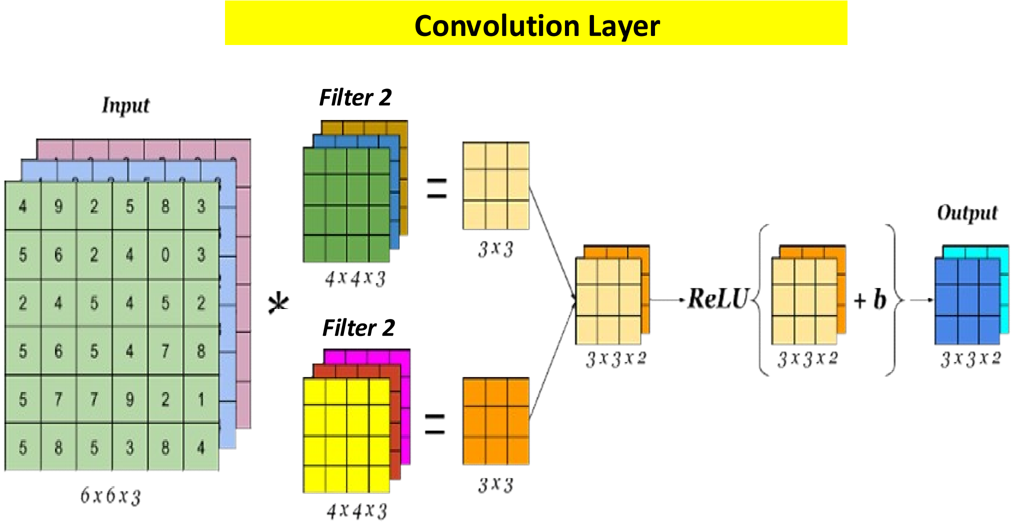


Convolution Layer

This layer is the core of the neural network. An input image is transformed for extracting the features, convolved through a kernel that has a small matrix having a height and length being smaller in size than the input image. Afterwards, the kernel slides all over in length and width across the input image, after which a dot matrix is calculated. To decrease and eradicate the non-linearity from the output ReLU, Tanh or any other activation functions are utilized.

Figure 2

Detailed Architecture of the Convolution Layer



4	9	2	5	8	3
5	6	2	4	0	3
2	4	5	4	5	2
5	6	5	4	7	8
5	7	7	9	2	1
5	8	5	3	8	4

Pooling Layer

The purpose of this layer is the reduction of the input image. This reduction adds more strength to the features and makes the computations fast. This layer utilizes kernel (filter) and stride. Pooling can be of different types like max pooling, and average pooling.

Fully Connected Layer

This layer has all the neurons that are connected with one another both in the successor and predecessor layers. In this layer, the input is multiplied by the weight matrix and then bias is added to it.

Activation Function

In a neural network, neurons play a vital role which taking the weighted sums of the inputs and passing the resultant scalar values to the function named as an activation function. The main purpose of the activation function is to determine whether the value of input remains the same or larger than the threshold value to activate the neuron. Whenever the value of the input is smaller than the threshold value, its output value will not be sent to the next layer as no neuron will be activated.

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \qquad (1)$$

Activation functions can be of various types and each performs well regardless of the problem. Here in our case, we have used ReLU, sigmoid and SoftMax.

ReLU is the most commonly and widely used activation function used in almost every deep learning network. The function works as follows:

$$\begin{cases} 0, if \;\; x < 0 \\ x, if \;\; x \geq 0 \end{cases} \qquad (2)$$

which means that the function will deliver the value back whenever the input is positive, otherwise it will return zero.

The S-shaped classification output displays a sigmoid activation function ensuing 0 or 1. It is described as follows;

$$A = \frac{1}{(1+e^{-x})} \qquad (3)$$

Similarly, a SoftMax activation function plots all the values of vectors into the probability vectors. This type of activation function forecasts spreading probability.

ML-DL SafetyNet Network

The proposed network ML-DL SafetyNet is based on deep learning and machine learning networks. The network comprises various layers. The output vectors of a fully connected layer are:

$$y^t = [y_1 + \; y_2 \; + \; y_3 \; + \ldots . + y_n] \qquad (4)$$

$$y = f(Wx \; + \; b)(1) \qquad (5)$$

$$x^t = [x_1 + x_2 + x + \ldots + x_m] \qquad (6)$$

$$b^t = [b_1 + b_2 + b_3 + \ldots + b_n] \qquad (7)$$

Where f, b and w are the activation function, input column vector and biases respectively.

The network comprises various layers having convolution, max pooling and ReLU layers. Activation functions play as the backbone of a neural network. These activation functions facilitate the neural network [23] with non-linearity as they assist the network in learning complex patterns in the system. Our ML-DL SafetyNet encompasses two portions i.e. deep learning and machine learning.

Deep Learning-based SafetyNet Network

Figure 4

Mel-Spectrogram for Spoof Audio

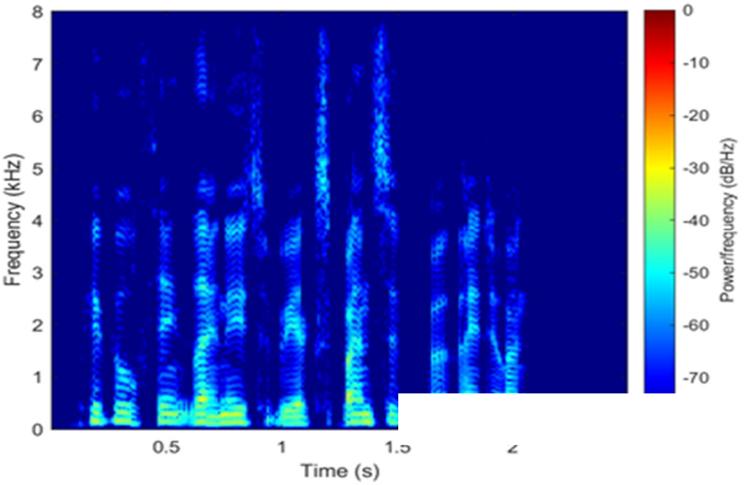
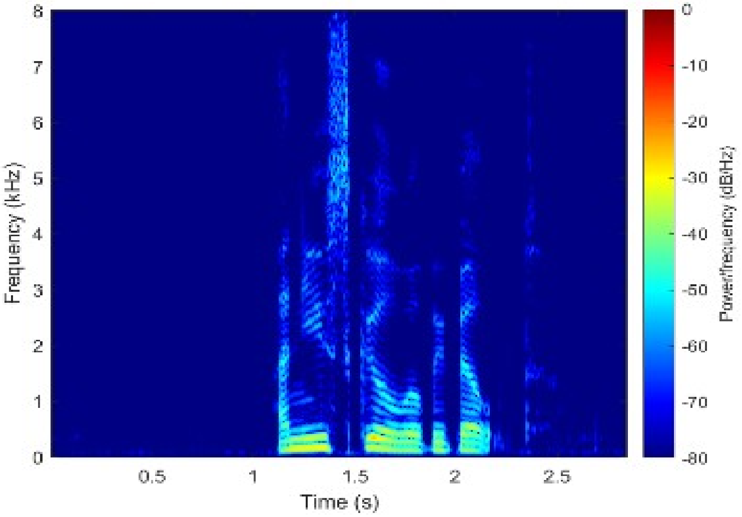


Figure 5

Mel-Spectrogram for Bonafide Audio



As the network has no provision of WAV audio files and exclusively accepts only the images as input, so initially we transformed all the audio WAV files into images of spectrograms to improve our system efficiency. After auditory records are converted into spectrograms, different layers like convolution, ReLU, sigmoid, fully connected, SoftMax and classout are performed on spectrograms. In the layer of image input, we have taken the audio as spectrograms. As Mel-Spectrogram images are converted, these images are then passed through the convolution layer for convolving into a smaller image by using a kernel or a filter. After the convolution, for the activation function, ReLU is utilized which avoids the exponential development in computation for a neural network. After this layer again, we have to pass our data over the convolution layer that benefits the system to absorb features upon variance scale on the image. The sigmoid function layer is then utilized which benefits to diminish the non-linearity and additionally, it ascertains the type of values to be passed and stopped as output. Furthermore, two fully connected layers are utilized for weights and biases to achieve the maximum desired results.

Figure 6

Deep Learning Network Detailed Layers Flow Chart

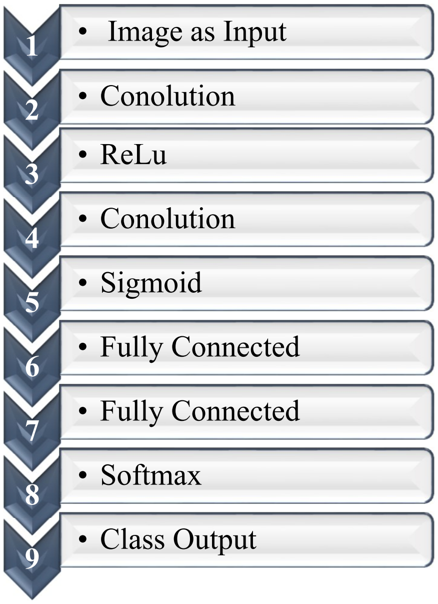
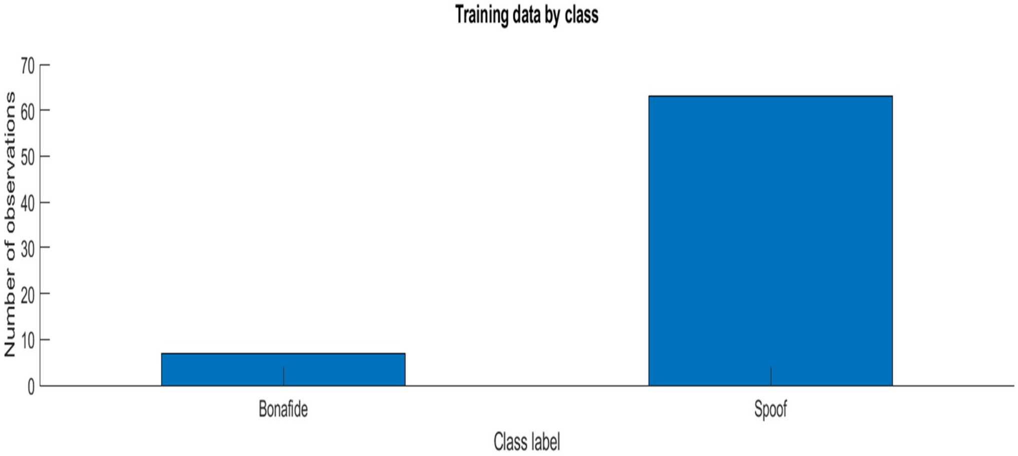


Figure 3

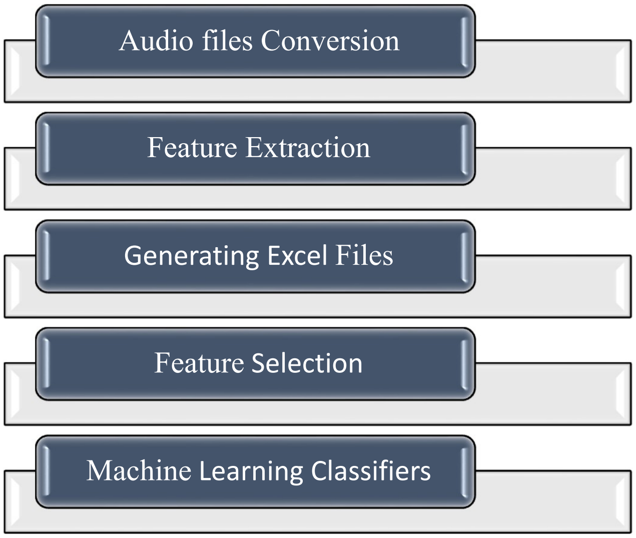
Training of Data by Class



Moreover, the SoftMax activation function generates the outputs of the network vectors into probability vectors. In the final stage classification output layer easily classifies and differentiates our results into desired classes on the basis of probability vectors. The crucial parameter to be considered is the learning rate in the training of deep learning and machine learning models. The learning rates considered in the ML-DL SafetyNet model are 0.01 and 0.0001. Likewise, we have used 10-fold cross-validation. Architectural details of the layers are described in table 2 below.

We navigated through a series of consecutive steps in the preceding deep learning approach to achieve the desired outcomes. In the same way, we have used another approach to attain our desired results by using machine learning. ASV spoof 2019 dataset comprises audio files in the format of .flac files. Our initial approach aims to convert audio data files into .wav files. After the conversion of .flac files to .wav files, a tool named JAudio tool for feature extraction was utilized. JAudio tool lacks the support of .flac files and accepts only .wav files. Accordingly, .flac files were converted to .wav files. In this way, we attained the features in the XML format and furthermore, features converted in Excel format. Likewise, feature selection is applied and by using different machine learning algorithms data is analyzed with accuracies, confusion matrices, ROC curves and coordinate plots. We applied various machine learning algorithms to attain the desired results. Among seven learners we achieved the best accuracy of 90% for support vector machine classifier. Therefore, in our approach, SVM is the best classifier to achieve good results.

Figure 4



Detailed Flowchart for Machine Learning Classifiers

Machine Learning-based SafetyNet Network

In the case of the machine learning model initially, the dataset was loaded comprising of audio files. Loading of these audio files created an audio Datastore (ADS) that can effectively load audio. Furthermore, the most important and necessary step in every model is feature extraction. For this purpose, our proposed model of machine learning achieved feature extraction. A variety of audio features were extracted from the loaded audio files of the ASV Spoof dataset. Mel-frequency Cepstral Coefficients, GammatoneCepstral Coefficients, flux, centroid, crest, decrease, entropy, flatness, kurtosis, roll-off point, skewness, slope, spread and energy are the features that are extracted. Moreover, the average of these extracted features was taken, an Excel file was generated and correlation analysis was carried out for these sets of features and achieved the correlation coefficients. When correlation analysis is performed then different machine learning algorithms are utilized and run to achieve the best result. After applying different machine learning algorithms like KNN, Gaussian Naive Bayes, Optimizable Tree, Logistic Regression, Linear Discriminant and Support Vector Machine. The results of these classifiers are shown in Fig 13.

Table 2

Architectural Details of Layers

S.No	Name	Type	Activation	Learnable
1	Image Input 227x227x1	Image Input	227x227x1	-
2	Conv_1 32 3x3 convolution with stride [1 1]	Convolution	227x227x32	Weights 3x3x1x32 Bias 1x1x32
3	Leaky ReLU_1 Leaky ReLU with a scale of 0.01	Leaky ReLU	227x227x32	-
4	Maxpool_1 5x5 max pooling with stride [1 1] and padding the same	Max Pooling	227x227x32	-
5	Conv_2 32 3x3 convolution with stride [1 1]	Convolution	227x227x32	Weights 3x3x1x32 Bias 1x1x32
6	Leaky ReLu_2 Leaky ReLU with a scale of 0.01	Leaky ReLU	227x227x32	-
7	Avgpool2d_1 5x5 average pooling with stride [1 1] and padding the same	Average Pooling	227x227x32	-
8	Conv_3 32 3x3 convolution with stride [1 1]	Convolution	227x227x32	Weights 3x3x1x32 Bias 1x1x32
9	Relu_1 ReLU	ReLU	227x227x32	-
10	Conv_4 32 3x3 convolution with stride [1 1]	Convolution	227x227x32	Weights 3x3x1x32 Bias 1x1x32
11	Leakyrelu_3 Leaky ReLU with scale of 0.01	Leaky ReLU	227x227x32	-
12	Maxpool_2 5x5 max pooling with stride [1 1] and padding the same	Max Pooling	227x227x32	-
13	FC_1 10 fully connected layer	Fully Connected	1x1x10	Weights 10x1648928 Bias 1x1x32
14	Leakyrelu_4 Leaky ReLU with a scale of 0.01	Leaky ReLU	227x227x32	-
15	FC_2 10 fully connected layer	Fully Connected	1x1x10	Weights 10x1648928 Bias 1x1x32
16	Relu_2 ReLU	ReLU	227x227x32	-
17	Conv_5 32 3x3 convolution with stride [1 1]	Convolution	227x227x32	Weights 3x3x1x32

S.No	Name	Type	Activation	Learnable
				Bias 1x1x32
18	Avgpool2d_2 5x5 average pooling with stride [1 1] and padding the same	Average Pooling	227x227x32	-
19	FC_3 10 fully connected layer	Fully Connected	1x1x10	Weights 10x1648928 Bias 1x1x32
20	Softmax Softmax	Softmax	1x1x10	-
21	Classoutput Classentropyex	Classification Output	1x1x10	-

The above table depicts the details of the architecture of layers used for our CNN model designed for the classification First of all the entry point of our CNN model is the image input as we converted our audio files into image files that can be easily compatible and supported by the proposed network model. 32 filters of size 3x3 were applied by the first convolution layer named Conv_1 on an input image. To enhance the competence and ability of the network to apprehend the complex patterns Leaky ReLU activation function was employed with a scale of 0.01. For maintaining the latitudinal proportions, a 5x5 max pooling layer was then performed through a stride of [1,1]. Likewise, Conv_1 another layer of 32 filters with 3x3 size was applied named Conv_2, by utilizing Leaky ReLU activation. To down sample the feature maps 5x5 average pooling was performed with stride [1,1] and padding Furthermore, a fully connected layer was introduced comprising 10 output nodes involving a substantial number of learnable parameters. In the end, the softmax function was applied to generate probability distribution across the classes and a classification output layer was utilized that performed a precise entropy-based loss function for training.

Results and Experiments

Experimental Set-Up

This section covers the details of the results which are carried out in the research. For the conduction of results, MATLAB software is used for the evaluation of the ASV Spoof 2019 dataset. The system used for research comprises of following specifications as shown in Table 3 below.

Table 3

Proposed Network - System Specifications

Hardware		
Laptop Specification:		
Processor	:	Core i7
RAM	:	16 GB
SSD	:	256 GB
Hard Drive	:	1 TB
GPU	:	2 GB
LCD	:	14´´
Software:		
A tool which is used for achieving the results and analysis of the dataset is MATLAB.		

Accuracy shows that a model is correct up to which level. When you want to find a specific class, accuracy tells how better this class is predicted.

Mathematically,

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (8)$$

Precision is basically the ratio of the predictions. Precision tells the ratio of estimates classified as confident which are appropriately classified to the number of estimates classified as confident whether they may be accurate or improper.

Mathematically,

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (9)$$

Recall is also the ratio of the predictions. Recall displays the relation between the number of confident examples that are appropriately projected and classified as confident to the total number of confident examples.

Mathematically,

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (10)$$

Dataset

ASV spoof 2019 dataset is trailed from the previous three sessions that were held during inter-speech 2013, 2015 and 2017. The first edition 2013 [Delgado2021] mainly concerned and targeted awareness about the spoofing threats. In the second edition of 2015 main target was to discriminate between real and fake speech by text-to-speech or voice alteration systems. Likewise, the next edition 2017 primarily targeted the detection of spoof attacks.

The latest edition of ASV spoof 2019 [Lorenzo2018]is the best at the moment for the ASV spoof 2015 dataset consists of TTS and VC-generated spoofing occurrences. Since remarkable progress occurred in the field of artificial intelligence, nowadays the tools and techniques which are used for the VC and TTS are so strongly developed that it is quite hard for someone to discriminate between false and genuine speech. Furthermore, as these tools provide much genuineness in speech, therefore, these threats are alarming and need to be addressed as soon as possible. Genuine speech was taken from 107 talkers among them 46 men and 61 women, having no noise effect in it. Various algorithms are utilized to create spoofed speeches from the genuine speech taken from 107 speakers. The dataset is segregated as follows:

Table 4

LA Spoofing System and Algorithms Details

Logical Access	Samples	Spoofing System	Algorithm Type	Data
Total Samples	25380	_	_	_
A01	3800	TTS	neural waveform model	Text
A02	3800	TTS	Vocoder	Text
A03	3800	TTS	Vocoder	Text
A04	3800	TTS	waveform concatenation	Text
A05	3800	VC	Vocoder	Speech
A06	3800	VC	spectral filtering	Speech
A07	3800	TTS	vocoder+GAN	Text
A08	3800	TTS	neural waveform	Text
A09	3800	TTS	Vocoder	Text
A10	3800	TTS	neural waveform	Text
A11	3800	TTS	griffin lim	Text
A12	3800	TTS	neural waveform	Text
A13	3800	TTS_VC	Waveform concatenation, waveform filtering	speech
A14	3800	TTS_VC	Vocoder	speech
A15	3800	TTS_VC	neural waveform	speech
A16	3800	TTS	waveform concatenation	text
A17	3800	VC	waveform filtering	speech
A18	3800	VC	Vocoder	speech
A19	3800	VC	spectral filtering	speech

Spoofed and genuine speeches were collected from the 20 speakers for the training of the dataset. For this purpose, some of the algorithms are used for speech conversion and some are used for speech synthesis as shown in table 5 below.

Table 5

Speech Conversion Algorithms Details

Task	Technique Used
Voice Conversion	▪ Neural Network ▪ Transfer Function
Speech Synthesis	▪ Neural network-based parametric speech synthesis using source filter vocoders ▪ Neural network based ▪ parametric speech synthesis using Wavenet ▪ Waveform concatenation

Table 5

ASVspoof 2019 LA and PA Dataset

PA Samples									
	Sample	Label				Type of Attack	Labels		Device
			a	b	c		a	b	
Training	54,000	Room Size (m^2)	2-5	5-10	10-20	Attacker to Talker Distance (cm)	10-50	Perfect	Perfect
Dev	33,534	T60 (ms)	50-200	200-600	600-1000	Attacker to Talker Distance (cm)	50-100	High	High
Eval	153,522	Talker to ASV Distance (cm)	10-50	50-100	100-150	Attacker to Talker Distance (cm)	>100	Low	Low
LA Samples									
		Spoof System		Input	Mechanism of Input		Generator for Sound Wave		
Training	25,380	A01, A02, A03, A04, A05, A06, A07, A08, A09, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19		Text Human Speech	TTS, RNN, NLP, WORLD, MFCC, Spectral filtering ASR, Conv and Bi Waveform Conc		Wave Net*, WORLD, Waveform Concatenation, OLA and Special Filters, Griffin-Lim, STRAIGHT, MFCC Vocoder.		
Dev	24,986								
Eval	71,933								

The main objective of ASV spoof creativity was to guard the programmed talker authentication from deceiving attacks [Kinnunen2017][27][28][29][Gomez2017] The below table highlights the LA and PA datasets for ASV spoof 2019.

Data for Experiments

The experimental approach for the research is constructed on training and testing of the LA dataset. LA dataset comprised 25,380 samples including both genuine and fake samples. Among these samples 2,590 samples are bonafide and 22,900 samples are spoofed. This research study falls under the interdisciplinary approach of a study converging data science, management science and linguistics. Audios contain linguistic data in the arrangement of verbal language. The present study has discovered automatic spoofing detection in the linguistic data of audio through deep learning.

During the training of the model, different limitations for the model are to be developed. Initially, the model was accomplished using a learning rate of 0.01 and after that, we also trained the model at the learning rate of 0.0001. In both cases of using different learning rates the accuracy of the trained model remained unaltered, which means the learning rate did change the model's accuracy. The limitations of the trained ML-DL SafetyNet model are as under:

Table 6

Limitations for the Trained ML-DL SafetyNet Model

Limitations	Value
Learning Rate	0.01 / 0.0001
Batch size	128
Confidence value	0.20
No. of epochs	30

Figure 5

Learning Rate: 0.0001 (Accuracy 90%) The Blue Curve Showing the Accuracy While the Orange Curve Showing Loss

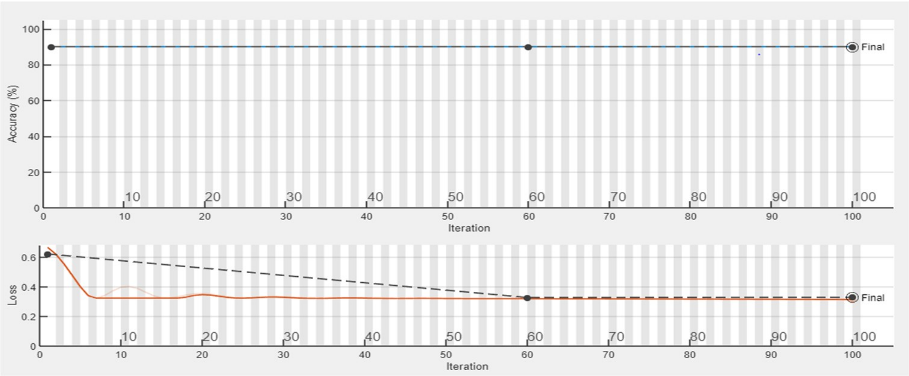
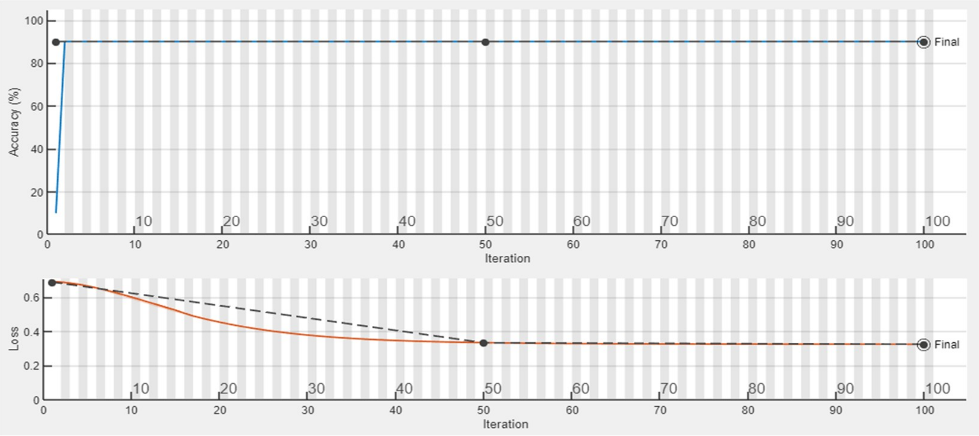


Figure 10

Learning Rate: 0.01 (Accuracy 90%) The Blue Curve Showing the Accuracy While the Orange Curve Showing Loss



Evaluation of Logical Access Violation's Performance

The main objective of the ML-DL SafetyNet model is to evaluate the effectiveness of the proposed model against the LA attacks. We have used classification learner algorithms by using machine learning and deep learning to differentiate between bonafide and spoof speeches. The table below indicates an equal minimum classification error and AUC in our proposed spoofing detection model. ML-DL SafetyNet model achieved different accuracies for different classification algorithms. Likewise, by using the deep learning ML-DL Safety Net model successfully extracted the features from the Mel-spectrograms. As far as our outcomes it is understood that the ML-DL SafetyNet model achieved better results.

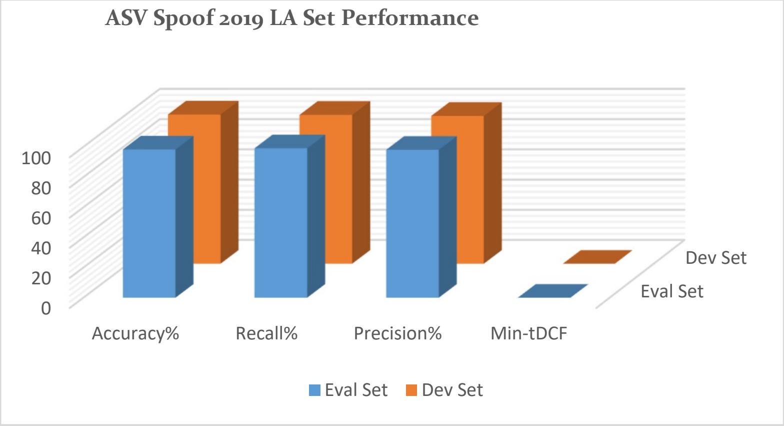
Table 8

Results of ASVspoof 2019 LA Corpus

Dataset	Accuracy%	Recall%	Precision%	Min-tDCF
Eval Set	98.3	99.1	98.1	0.003
Dev Set	98.8	98.6	97.9	0.007

Figure 11

ASVspoof 2019 LA Dataset Performance Plot



Performance Comparison of Physical Access Attack

The foremost objective of the research is to dig out the usefulness of spoofing recognition for physical access attacks. For such a reason audio samples are taken from the PA set and their Mel spectrograms are created after the spectrogram's generation, these spectrograms are separated based on real and fake classes. The results indicate an EER of 0.62% and 3.4% for eval and dev sets respectively. Similarly, results indicate min-tDCF of 0.04 and 0.09 for eval and dev sets respectively. As far as accuracies are concerned, we have accomplished better results from the earlier research models as we have achieved 99.5% and 97.4% accuracies for eval and dev sets. Performance plots and results are shown below.

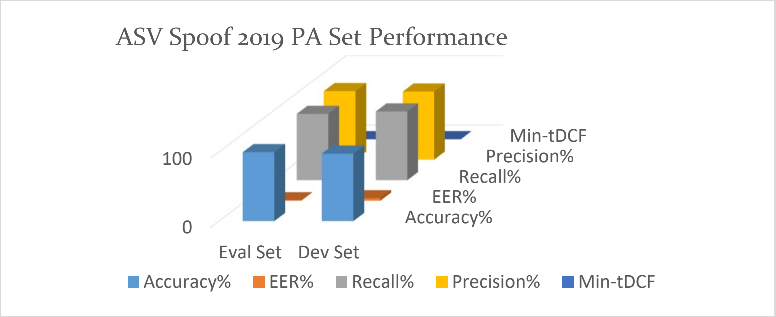
Table 9

Dataset	Accuracy%	EER%	Recall%	Precision%	Min-tDCF
Eval Set	99.5	0.62	95.84	99.2	0.04
Dev Set	97.4	3.4	99.34	98.6	0.09

ASVspoof 2019 Dataset PA Corpus Results

Figure 12

ASVspoof 2019 PA Dataset Performance Plot



Comparison of Various Machine Learning Classifiers

Our research is focused on audio spoofing detection which is more hazardous than video deep fakes because most of our communication is based on audio like audio phone calls, voice recordings etc. For such a reason, it is the essential need of moment to recognize fake audios. We calculated the results of our problem by using deep learning and machine learning techniques. By using deep learning, we performed the 10-fold cross-validation and used different learning rates, but the output did not change by changing the learning rate. We achieved an accuracy of 90% by using a deep learning app. Similarly, we used different classification learners on the same data to achieve the results. We performed seven classifiers namely KNN, SVM, fine tree, naïve Bayes, logistic regression, linear discriminant and optimizable discriminant. Among these classifiers support vector machine (SVM) performed well with an accuracy of 90%. The results of the different classifiers are as under:

Table 10

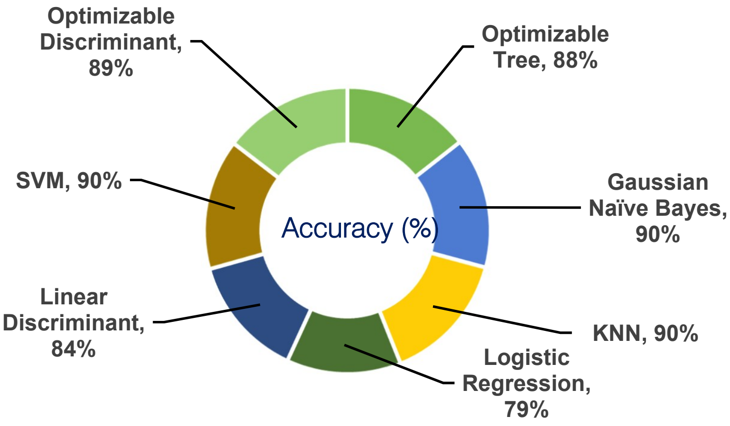
Accuracies for Various ML Classifiers

S. No	Algorithm Name	Accuracy (%)
1	Optimizable Tree	88
2	Gaussian Naïve Bayes	90
3	K-Nearest Neighbor (KNN)	90
4	Logistic Regression	79
5	Linear Discriminant	84
6	Support Vector Machine (SVM)	90
7	Optimizable Discriminant	89

The graphical representation of the classifiers is shown in Figure 13 below.

Figure 6

Graphical Representation of ML Classifiers Performance



Recognition of Voice Copying Algorithms

The principal ambition of the research is to figure out which technique and algorithm applied is to be best utilized in future for such types of problems. Basically, six different kinds of algorithms were used in the ASV spoof 2019 dataset as per the LA group is concerned (i.e. A01 to A06)

In the experimental planning 25,380 samples were used for the determination of model training and to check whether the model is best and how the model will behave for the samples used. 24,986 samples were collected for the development set to test the model. The algorithms (A01 to A06) used in this experimentation accomplished several enhanced entity relationships (i.e. 0.5, 2.0, 1.09, 1.2, 1.02 and 1.6 % respectively).

Artificial Voice and Voice Conversion Performance Review

This experimental setup is focused on the investigation of the results that spoofing is well identified for which of the techniques either text-to-speech or the voice cloning technique. We have implemented the Mel-spectrograms with various different characteristics for the training of our dataset samples including TTS and VC samples. Experimentation comprises the various Voice Cloning (VC) systems including A05 and A06 for the generation of the spoofing trials. Similarly, experimentation also includes text-to-speech (TTS) systems (A01 to A04) for the generation of various spoofing trials for our model. Both these systems' main objective was to generate the spoof trials for the LA dataset for our model to be used for training purposes. Moreover, the LA dataset includes the evaluation set which contains other different 13 algorithms (i.e. A07 to A12, A16 TTS spoofing systems, A17 to A19 VC spoofing systems and A13 to A15 VC-TTS spoofing systems). Experimental results are shown below.

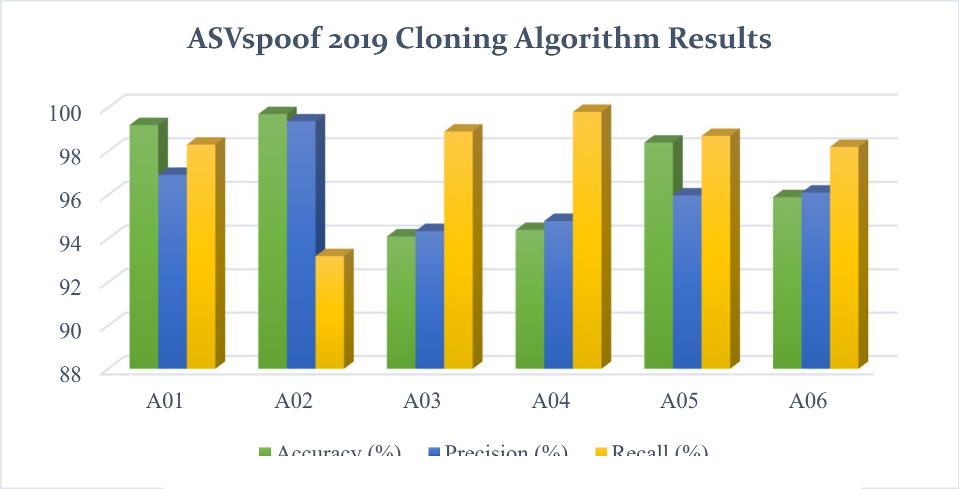
Table 11

ASVspoof 2019 Cloning Algorithms Results

Algorithm	Accuracy (%)	Precision (%)	Recall (%)
A01	99.2	96.92	98.3
A02	99.7	99.37	93.2
A03	94.1	94.35	98.9
A04	94.4	94.81	99.8
A05	98.4	95.98	98.7
A06	95.9	96.11	98.2

Figure 14

ASVspoof 2019 Dataset Cloning Algorithms Performance Plots



TTS spoof trials generated and real from the ASVspoof 2019 dataset were used for the training of the model and at the same time the trials from the evaluation set of TTS were used to test the proposed model. Enhanced entity-relationship (EER) achieved from the model is 0.63% and min-tDCF is 0.0159.

Table 12

Artificial Voice and Speech Development Results

Spoofing Category	Min-tDCF	EER (%)	Accuracy (%)	Precision (%)	Recall (%)
Voice Cloning (VC)	0.39	18.6	83.10	98.21	76.95
Text to Speech (TTS)	0.04	0.49	98.99	99.33	98.96
Overall LA	0.03	0.08	99.45	99.41	99.29

Disparity of Performance to Prevailing Techniques

The experiment compares the speech spoofing detector to other voice spoofing detection techniques. We conducted a comparison analysis with the models listed below in the table to demonstrate the viability of the ML-DL SafetyNetmodel, which is an improved ANN-based classifier for good detection of flaws in playback samples, cloning algorithms, art facts and authentic samples' dynamic speech qualities depending on their vocal tracts. The proposed and current approaches' performance in the context of EER and min-tCDF results on PA and LA datasets of ASV spoof 2019 are shown.

Table 13

Voice Spoofing Detection Recent Research Comparison

System	LA Evaluation Set		PA Evaluation Set	
	% of EER	Min-tDCF	% of EER	Min-tDCF
Baseline: GMM [31]	2.71	0.0663	8.09	0.2116
MFCC [31]	16.80	0.3945	-	-

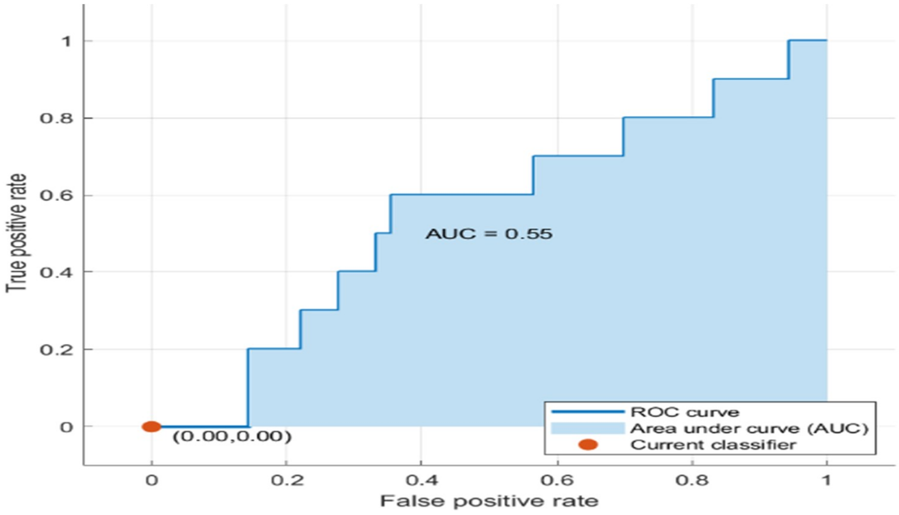
System	LA Evaluation Set		PA Evaluation Set	
	% of EER	Min-tDCF	% of EER	Min-tDCF
CQCC [31]	8.82	0.2076	12.06	0.2982
Baseline: GMM [32]	0.43	0.0123	9.57	0.2366
LFCC [32]	2.71	0.0663	8.09	0.2116
CQCC [32]	0.94	0.08	0.43	0.03
Baseline: GMM [33]	2.64	0.0755	5.43	0.1465
LFCC [33]	8.09	0.2116	13.54	0.3017
CQCC [33]	9.57	0.2366	11.04	0.2454
Baseline: GMM [34]	10.62	0.2401	5.58	0.1518
LFCC [34]	0.28	0.0062	4.79	0.1314
CQCC [34]	0.43	0.0123	9.87	0.1953
Baseline: GMM [35]	-	-	-	-
LFCC [35]	11.04	0.2454	0.43	0.0123
CQCC [35]	9.87	0.1953	9.57	0.2366
Baseline: GMM [36]	5.06	0.1562	-	-
LFCC [36]	4.04	0.1655	-	-
CQCC [36]	2.64	0.1331	-	-
Baseline: Parallel DDWS [37]	2.63	-	0.47	-
Sequential DDWS [37]	2.08	-	0.63	-
BC Res-Max [37]	2.59	-	0.49	-
CGCNN: VAE log-CQT + log CQT [38]	1.84	0.056	0.35	0.0092
CGCNN: Phase + log CQT [38]	1.09	0.034	0.31	0.0078
ResNet18: Phase + log CQT [38]	1.53	0.051	1.16	0.0350
Baseline: CLS-LBP + LSTM [39]	0.06	0.0017	0.58	0.0160
CQCC [39]	1.18	0.0520	11.5	0.2457
Ours:DLDet	0.052	0.0028	0.41	0.0243

Receiver Operating Curves (ROC)

Receiver operating curves basically assist in designing two types of factors i.e. True Positive (TP) and False Positive (FP). ROC curves for all the classifiers are plotted. Basically, these curves help to express the performance of the designed classifier models. ROC curves along with the area under the curves (AUC) are also plotted in the same window. The plots are as under:

Figure 7

ROC Curve Plot for Proposed Model



Conclusion

The proposed ML-DL SafetyNet model is structured into two sections, the first section powers deep learning techniques, while the second section uses machine learning techniques. In the first section of the ML-DL SafetyNet model, ASV spoof 2019 dataset audio files of logical access (LA) are utilized and converted to image spectrograms that are supported by MATLAB. Afterwards, the ML-DL SafetyNet model is trained using different learning rates. The proposed model achieved an accuracy of about 90% by using a deep learning approach. In the case of the second approach of ML-DL SafetyNet feature extraction and feature selection are performed by using various machine learning classifiers. Likewise, the same ASV spoof 2019 dataset was used for machine learning classifiers as in the preceding section of deep learning and performed the tasks over seven classifiers. Among all the classifiers, the support vector machine (SVM) performed best with an accuracy of 90%. Our relative analysis of the existing models proposes that our ML-DL SafetyNet model outperforms in perceiving various sorts of speech spoofing, including TTS, replay attacks and cloning-based attacks. It is noteworthy, that our model established excellent results on ASV spoof 2019. We can conclude that model ML-DL SafetyNet is a strong deceiving indicator, authenticated through the usefulness in cross justification over the ASVspoof 2019 evaluation set. In future, our ambition is to encompass cross-validation to different speech-deceiving datasets and additionally enhance the model's performance.

References

Almutairi, Z.,& H. Elgibreen(2023). "Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning," *IEEE Access, 1,*https://doi.org/10.1109/ACCESS.2023.3286864.

Google Scholar Worldcat Fulltext

Alzantot, M., Wang, Z.,&B. Srivastava, (2019)"Deep residual neural networks for audio spoofing detection," in Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 1078–1082. https://doi.org/10.21437/Interspeech.2019-3174.

Google Scholar Worldcat Fulltext

Balamurali, B. T. Lin, K. E. S. Lui, J. M. Chen, & D. (2019). Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access, 7,* 84229–84241, https://doi.org/10.1109/ACCESS.2019.2923806.

Google Scholar Worldcat Fulltext

Columbia,B.(2021). "A Capsule Network Based Approach for Detection of Audio Spoofing Attacks 1. Key Lab of Information Security, School of Computer Science and Engineering, Sun Yat-Sen University, 2. Alibaba Group, Hangzhou, China," 6359–6363,

Google Scholar Worldcat Fulltext

Delgado,H. et al., (2021). "ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," http://arxiv.org/abs/2109.00535

Google Scholar Worldcat Fulltext

Dua, M. C. Jain, & Kumar, S.(2022). "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems," J. Ambient Intell. Humaniz.*Comput, 13*(4), 1985–2000, https://doi.org/10.1007/s12652-021-02960-0.

Google Scholar Worldcat Fulltext

Gao, Y.Vuong, M. Elyasi, G. Bharaj, & Singh, R.(2021). "Generalized Spoofing Detection Inspired from Audio Generation Artifacts," http://arxiv.org/abs/2104.04111

Google Scholar Worldcat Fulltext

Gomez-alanisA. et al. (2017). "On Joint Optimization of Automatic Speaker Verification and Anti-spoofing in the Embedding Space," *i,* 1–15.

Google Scholar Worldcat Fulltext

Hamza, A. et al. (2022). "Deepfake Audio Detection via MFCC features using Machine Learning," *IEEE Access, 10*, 134018–134028, https://doi.org/10.1109/ACCESS.2022.3231480.

Google Scholar Worldcat Fulltext

Ismail, A. M. Elpeltagy, M. S. Zaki., & K. Eldahshan, "A New Deep Learning-Based Methodology for Video Deepfake," 1–15.

Google Scholar Worldcat Fulltext

Kinnunen, T. et al. (2017). "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 2–6. https://doi.org/10.21437/Interspeech.2017-1111.

Google Scholar Worldcat Fulltext

Kinnunen, T. M. Todisco, N. Evans, J. Yamagishi., & K. A. Lee, (2017). "The ASVspoof 2017 Challenge : Assessing the Limits of Replay Spoofing Attack Detection National Institute of Informatics, Japan," *i*, 2–6.

Google Scholar Worldcat Fulltext

Lavrentyeva, G. S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev., &Shchemelinin, V. (2017). "Audio replay attack detection with deep learning frameworks," in Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 82–86. https://doi.org/10.21437/Interspeech.2017-360.

Google Scholar Worldcat Fulltext

Lorenzo-Trueba,J. et al.(2018). "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods," http://arxiv.org/abs/1804.04262

Google Scholar Worldcat Fulltext

Mcuba, M. A. Singh, R. A. Ikuesan., & H. Venter, (2022). "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital

Investigation," Procedia Comput. *Sci., 219,* 211–219. https://doi.org/10.1016/j.procs.2023.01.283.

Google Scholar Worldcat Fulltext

Todisco, M. et al.(2019). "ASVSpoof Future horizons in spoofed and fake audio detection," Proc. Annu. Conf. Int. Speech Commun. Assoc. *Interspeech,* 1008–1012, https://doi.org/10.21437/Interspeech.2019-2249.

Google Scholar Worldcat Fulltext

Wang, Z. S. Cui, X. Kang, W. Sun., & Z. Li, (2020). "Densely Connected Convolutional Network for Audio Spoofing Detection," 1352–1360.

Google Scholar Worldcat Fulltext

Wenger, E. M. Bronckers, C. Cianfarani, J. Cryan, A. Sha., & B. Y. Zhao, (2021). "Hello, It's Me": Deep Learning-based Speech Synthesis A acks in the Real World," 235–251.

Google Scholar Worldcat Fulltext

Yang, Y. et al.(2019). "The SJTU Robust Anti-spoofing System for the ASVspoof 2019 Challenge," 1038–1042, https://doi.org/10.21437/Interspeech.2019-2170

Google Scholar Worldcat Fulltext

Yi, J. et al.(2021). "Half-truth: A partially fake audio detection dataset," Proc. Annu. Conf. Int. Speech Commun. Assoc. *INTERSPEECH, 4,* 2683–2687, https://doi.org/10.21437/Interspeech.2021-930.

Google Scholar Worldcat Fulltext

Yi, J et al.(2022). "ADD 2022: the First Audio Deep Synthesis Detection Challenge," http://arxiv.org/abs/2202.08433

Google Scholar Worldcat Fulltext

Yu, Y. et al.(2020). "RMAF : Relu-Memristor-Like Activation Function for Deep Learning," *IEEE Access, 8,* 72727–72741, https://doi.org/10.1109/ACCESS.2020.2987829.

Google Scholar Worldcat Fulltext